BEYOND FIELD F1: CONCEPT-BINDING ACCURACY AND THE HINDSIGHT BENCHMARK FOR DOCUMENT AI

Anonymous authors
Paper under double-blind review

ABSTRACT

Document AI systems often extract text correctly but assign it to the wrong semantic field, for example, reading "$4,523.67" correctly from a paystub but labeling it *net pay* instead of *gross pay*. Value-only matching, which we term Value-Only Recall (Rec_{val}), credits a ground-truth value if it appears anywhere in the output, irrespective of its concept key, and therefore misses this failure. We introduce *Concept-Binding Accuracy (CBA)*, a metric family that decomposes joint key-value error into extraction and binding components, and *Hindsight*, a benchmark of nine datasets across five document domains with hierarchical concept ontologies. For Claude Haiku 4.5 and GPT-4o-mini, Rec_{val} overstates joint key-value accuracy by up to 33.3 percentage points: on commercial contracts, GPT-4o-mini reaches 100% Rec_{val} but only 66.7% CBA, with 250 misbindings across 50 documents. Even fixed-layout IRS W-2 forms yield 194 (Haiku) and 115 (GPT-4o-mini) misbindings when 44 same-type fields compete, whereas low-ambiguity domains show near-zero gaps. We identify two drivers, *concept density* and *semantic overlap*, and confirm both as independent causal effects in a controlled 5×4 factorial study on W-2: overlap alone moves binding accuracy by 25 percentage points, and density alone by up to 16 percentage points. An open-weights model (Qwen2-VL-7B) shows the same pattern, with markedly worse binding on the highest-ambiguity domain (CUAD) and comparable binding on low-ambiguity ones. We release all ontologies, scoring code, and experiment notebooks.

1 INTRODUCTION

Modern Document AI pipelines, from OCR and layout analysis to structured extraction, achieve near-human text reading accuracy (Tjong Kim Sang & De Meulder, 2003; Shi et al., 2017). Schema-aware key-value F1 scores joint (concept, value) tuples and therefore does penalize a value assigned to the wrong field. In practice, however, extraction pipelines are often evaluated with a weaker, value-only baseline that credits a ground-truth value if it appears *anywhere* in the output, irrespective of the concept key attached to it. We call this baseline **Value-Only Recall** (Rec_{val}). It conflates two capabilities: *value extraction* (reading the text correctly) and *concept binding* (assigning the value to the correct field). CBA is designed to diagnose this conflation, not to replace schema-aware F1.

When a document contains several fields of the same type, such as dollar amounts on a paystub, dates in a contract, or near-synonym legal clauses, a model can extract every value perfectly yet assign them to the wrong slots. We call this failure *concept misbinding*. GPT-4o-mini achieves perfect Rec_{val} on commercial contracts while producing 250 misbindings, a 33.3-percentage-point gap to CBA. On IRS W-2 forms with 44 labeled boxes, Claude Haiku 4.5 makes 194 misbindings and GPT-4o-mini 115. In automated pipelines, a misbinding passes every type and format check and propagates silently downstream. Our contributions are:

- *Metric and protocol:* the CBA metric family, with strict and soft (partial-credit) variants, and a value-first bipartite matching protocol that separates binding errors from extraction failures (Section 3).
- *Benchmark:* Hindsight, which couples nine datasets across five domains with hierarchical concept ontologies to diagnose within- and cross-family confusion (Section 3.4).
- *Failure analysis:* Rec_{val} materially overstates joint key-value accuracy, and CBA decomposes the error into extraction and binding components (Section 4.1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- *Drivers of misbinding:* misbinding concentrates where concept density and semantic overlap are high, and fixed layout does not remove it. A controlled factorial study (Section 4.5) isolates both drivers, and a prompt-format ablation (Section 4.4) shows that schema-constrained prompting is only a domain-dependent mitigation.

2 Related Work

Document extraction benchmarks. Standard information-extraction evaluation computes precision, recall, and F1 over matched field–value pairs (Tjong Kim Sang & De Meulder, 2003; Shi et al., 2017; Xu et al., 2020). FUNSD and VRDU use extraction-style metrics (Jaume et al., 2019; Wang et al., 2023). SROIE scores a few well-separated receipt fields by exact match (Huang et al., 2019), DocVQA scores each field independently as a question (Mathew et al., 2021), CORD uses structure-aware scoring (Park et al., 2019), and CUAD uses span-overlap retrieval metrics (Hendrycks et al., 2021). DocILE, OmniDocBench, and SCORE scale up structured extraction and perception fidelity (Simsa et al., 2023; Ouyang et al., 2025; Li et al., 2025). None isolates the case in which the correct value is attached to the wrong concept.

Structured outputs and attribute extraction. ExtractBench, VAREX, and LLMStructBench test whether models produce valid JSON under large schemas (Ferguson et al., 2026; Barzelay et al., 2026; Tenckhoff et al., 2026). Attribute-value extraction either assumes the attribute is given (Zheng et al., 2018) or scores joint pairs by exact match (Roy et al., 2022), and LLM information-extraction benchmarks report aggregate F1 (Han et al., 2023). All of these scores aggregate extraction and assignment. PARSE shows that ambiguous or overlapping schema keys induce extraction errors (Shrimal et al., 2025), a schema-design view of the drivers we measure.

Type confusion and slot filling. Fine-grained entity typing studies which types are confusable across a corpus-wide taxonomy (Ling & Weld, 2012; Gillick et al., 2016), and TAC-KBP slot filling disambiguates slots across knowledge bases (Ji et al., 2011). CBA asks the within-document analog: was a value bound to the correct *document-local* concept? Appendix A gives an extended discussion.

3 Concept-Binding Accuracy and the Hindsight Benchmark

3.1 A Three-Metric Framework

Let $GT = \{(c_i, v_i)\}_{i=1}^{N}$ be the ground-truth concept-value pairs for a document (or, summed additively, a corpus), where $c_i \in \mathcal{C}$ is a concept from the domain ontology and v_i is the ground-truth text value. A model produces predictions $P = \{(c'_j, v'_j)\}_{j=1}^{M}$. All value matching is performed under a domain-specific normalization $\mathrm{norm}(\cdot)$, using the value-first matching protocol of Section 3.3.

Definition 1 (Concept Misbinding)**.** A prediction $(c'_j, v'_j) \in P$ is a *misbinding* of $(c_i, v_i) \in GT$ if $\mathrm{norm}(v'_j) = \mathrm{norm}(v_i)$ but $c'_j \neq c_i$.

We use three metrics with explicit, distinct denominators. **Value-Only Recall** is the fraction of ground-truth values recovered anywhere in P, irrespective of concept key:

$$Rec_{val} = \frac{\left|\{(c_i, v_i) \in GT : \exists (c'_j, v'_j) \in P \text{ s.t. } \mathrm{norm}(v'_j) = \mathrm{norm}(v_i)\}\right|}{|GT|}. \tag{1}$$

Conditional Concept-Binding Accuracy is the fraction of *recovered* values (the set counted in the numerator of Equation 1) that were also assigned to the correct concept key:

$$CBA_{cond} = \frac{\left|\{(c_i, v_i) \in GT : \exists (c'_j, v'_j) \in P \text{ s.t. } \mathrm{norm}(v'_j) = \mathrm{norm}(v_i) \wedge c'_j = c_i\}\right|}{\left|\{(c_i, v_i) \in GT : \exists (c'_j, v'_j) \in P \text{ s.t. } \mathrm{norm}(v'_j) = \mathrm{norm}(v_i)\}\right|}. \tag{2}$$

Joint Key-Value Accuracy requires both the value and the key to be correct, over all of GT:

$$\begin{aligned} ACC_{joint} &= Rec_{val} \times CBA_{cond} \\ &= \frac{\left|\{(c_i, v_i) \in GT : \exists (c'_j, v'_j) \in P \text{ s.t. } \mathrm{norm}(v'_j) = \mathrm{norm}(v_i) \wedge c'_j = c_i\}\right|}{|GT|}. \end{aligned} \tag{3}$$

ACC_{joint} is the recall term of schema-aware key-value F1 (Section 3.2). Because CBA_{cond} conditions on recovery, it can exceed Rec_{val}: on GST, $Rec_{val} \approx 30\%$ while $CBA_{cond} \approx 97$–99%.

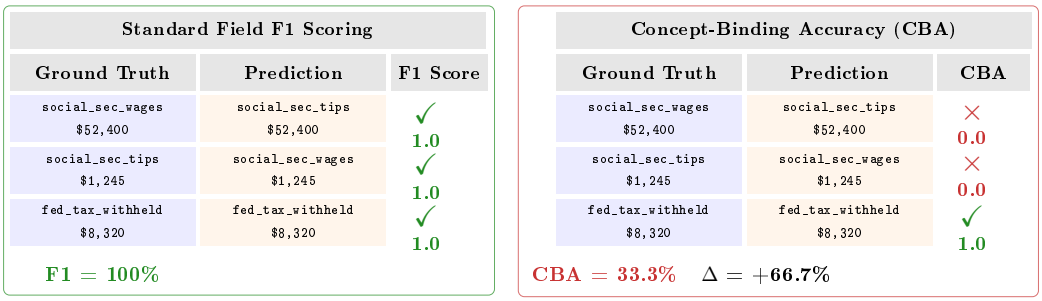


Figure 1: Comparison of Value-Only Recall (Rec_{val}) and Concept-Binding Accuracy (CBA) on an example where values are extracted correctly but assigned to the wrong concepts.

Strict and soft scoring. The numerator of CBA_{cond} is populated by a per-pair score for a ground-truth concept c_i matched to a predicted concept c'_j under an established value match. CBA-strict, which we report as CBA_{cond} throughout, assigns binary credit; CBA-soft gives partial credit when the two concepts share an ontology family (Section 3.4):

$$\text{CBA-strict}(c_i, c'_j) = \mathbf{1}[c_i = c'_j], \tag{4}$$

$$\text{CBA-soft}(c_i, c'_j) = \begin{cases} 1.0 & \text{if } c_i = c'_j \\ 0.5 & \text{if } c_i \neq c'_j \text{ and family}(c_i) = \text{family}(c'_j) \\ 0.0 & \text{otherwise.} \end{cases} \tag{5}$$

CBA-soft is an ontology-relative diagnostic, not a calibrated severity score (Appendix L.1).

Misbinding delta. We define Δ relative to ACC_{joint}:

$$\Delta = Rec_{val} - ACC_{joint} = Rec_{val} \times (1 - CBA_{cond}). \tag{6}$$

Δ is the fraction of *all* ground-truth items whose value was recovered but lost to misbinding. $\Delta \geq 0$ always, with $\Delta = 0$ exactly when $CBA_{cond} = 1$ or $Rec_{val} = 0$ (CBA_{cond} is then undefined, but $ACC_{joint} = 0$ and hence $\Delta = 0$ by its first form). The naive difference $Rec_{val} - CBA_{cond}$, by contrast, can be negative even when misbinding occurs, because its terms have different denominators. When $Rec_{val} = 1$ (LEDGAR and CUAD), $\Delta = 1 - CBA_{cond}$.

Worked example. Figure 1 shows a W-2 form with three fields whose values are all extracted, but with two of them swapped. $Rec_{val} = 3/3 = 100\%$, while $CBA_{cond} = ACC_{joint} = 1/3 = 33.3\%$, so $\Delta = +66.7\%$: two-thirds of the "correct" extractions are misbound. CBA-soft gives $(0.5 + 0.5 + 1.0)/3 = 66.7\%$, because the swap stays within the wages family.

3.2 Relationship to Schema-Aware Key-Value F1

Schema-aware key-value F1 scores joint tuples, so a misbinding counts as one false negative and one false positive. It therefore *does* penalize misbinding; Rec_{val} does not. What Rec_{val} and CBA add is a *decomposition* of joint error into value-reading failure ($1 - Rec_{val}$) and assignment failure among recovered values ($1 - CBA_{cond}$). A joint score of 66.7% on CUAD is consistent with many failure profiles. The decomposition ($Rec_{val} = 100\%$, $CBA_{cond} = 66.7\%$) shows that the failure is entirely assignment, which calls for better prompting rather than better OCR. Appendix B compares the metric families (Table 6).

3.3 Value-First Scoring Protocol

Ground-truth items and predictions form the two sides of a bipartite graph (Algorithm 1). Edge costs prefer a value-and-key match (0) over a value-only match (0.5, a misbinding) over no value match (1). `scipy.optimize.linear_sum_assignment` then returns the cost-minimizing one-to-one assignment, which avoids arbitrary tie-breaking when a normalized value occurs more than once. Ground-truth items with no value match do not count toward Rec_{val} and are excluded from the denominator of CBA_{cond}. The step-by-step rules are in Appendix B.

Algorithm 1 Value-First Bipartite Hungarian Matching

Require: Ground truth $GT = \{(c_i, v_i)\}_{i=1}^{N}$, predictions $P = \{(c'_j, v'_j)\}_{j=1}^{M}$
Ensure: Rec_{val}, CBA_{cond}, ACC_{joint}, misbinding count

1: **Step 1 (Normalize):** $\tilde{v}_i \leftarrow \text{norm}(v_i)$ for all i; $\tilde{v}'_j \leftarrow \text{norm}(v'_j)$ for all j
2: **Step 2 (Candidate graph):** Build bipartite graph B on (GT, P) with an edge (i, j) for every pair
3: **Step 3 (Cost matrix):** Construct $C \in \mathbb{R}^{N \times M}$:
$C_{i,j} = 0.0$ if $\tilde{v}_i = \tilde{v}'_j$ and $c_i = c'_j$ ▷ value and concept both correct
$C_{i,j} = 0.5$ if $\tilde{v}_i = \tilde{v}'_j$ and $c_i \neq c'_j$ ▷ value correct, misbinding
$C_{i,j} = 1.0$ otherwise ▷ no value match
4: **Step 4 (Optimal assignment):** $(i^*, j^*) \leftarrow$ LINEAR_SUM_ASSIGNMENT(C) ▷ `scipy.optimize`
5: **Step 5 (Classify):** For each assigned pair (i^*, j^*): classify as joint match ($C = 0$), misbinding ($C = 0.5$), or miss ($C = 1.0$); accumulate into Rec_{val}, CBA_{cond}, ACC_{joint} (Equations 1–3), and the misbinding count

Table 1: Concept ontologies across Hindsight domains and each domain's position in the density × overlap design space (Den./Ovl.: L = low, M = moderate, H = high; see Table 7).

Domain	Concepts	Families	Example Families	Den./Ovl.
Paystubs	14	5	Compensation, deductions, identifiers	M / L
SROIE	4	3	Amounts, identifiers, metadata	L / L
CORD	13	3	Amounts, payments, counts	M / L
W-2	44	8	Wages, taxes, identifiers, state_fields	H / H
LEDGAR	15	6	Termination, compensation, IP, liability	M / H
CUAD	15	5	Liability, temporal, exclusivity, assignment	M / H
VRDU Reg.	6	4	Names, signer_details, identifiers, dates	M / L
VRDU Ad-buy	9	4	Organisations, media, dates, financial	M / L
GST Invoices	13	4	Tax_components, amounts, dates, identifiers	M / L

3.4 Concept Ontologies and Benchmark Design

Each domain has a hierarchical ontology that groups concepts into semantic families, such as all wage fields or all tax fields. The families enable CBA-soft and confusion diagnostics (Table 1). Domains were chosen for four criteria: **C1** concept multiplicity, **C2** semantic proximity, **C3** low-ambiguity baselines to test specificity, and **C4** domain diversity. Together they span the concept-density × semantic-overlap design space (Table 1, last column; see also Tables 7 and 8 and Appendix E). The authors built the families from shared surface type, co-occurrence, and plausible confusability, and finalized them before any model was evaluated. Because this involves researcher judgment, we treat CBA-soft as ontology-relative (Appendix B.3).

3.5 Models and Extraction Setup

We evaluate Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) and GPT-4o-mini (`gpt-4o-mini-2024-07-18`), both cost-efficient production models, at T=0 with `random.seed(42)` sampling. The prompt format follows each domain's design. CUAD, LEDGAR, and Paystubs request a free JSON array of `{label, value, concept}` triples, so the model must name each concept. W-2, SROIE, and GST request a JSON object with fixed concept keys. Section 4.4 ablates this choice. Image domains use vision endpoints. LEDGAR and CUAD pass the document text, and Paystubs pass the Tesseract OCR text. Prompts, rate limits, and compute are described in Appendices F and H.

Table 2: Aggregated results across all benchmark domains, computed with the Hungarian-matching implementation (Algorithm 1), reporting Value-Only Recall (Rec_{val}), Conditional Concept-Binding Accuracy (CBA_{cond}), Joint Key-Value Accuracy ($ACC_{joint} = Rec_{val} \times CBA_{cond}$), $\Delta = Rec_{val} - ACC_{joint}$, and the number of concept misbindings for Claude Haiku 4.5 and GPT-4o-mini, with 95% bootstrap confidence intervals (10,000 paired document-level resamples) in brackets. $\Delta \geq 0$ for every row: it is the fraction of ground-truth items whose value was recovered but lost to misbinding, and cannot be negative under this formula (Section 3.1). CORD reflects its LLM-vision track only (Appendix E).

Dataset	Model	Rec_{val} (%)	CBA_{cond} (%)	ACC_{joint} (%)	Δ (%)	Misbindings
Paystubs	Haiku	71.1 [65.2, 76.9]	99.0 [97.3, 100.0]	70.4	0.7 [0.0, 1.9]	5
	GPT-4o-mini	62.5 [56.7, 68.0]	98.6 [97.2, 99.8]	61.6	0.9 [0.1, 1.7]	6
SROIE	Haiku	59.5 [54.5, 64.5]	99.2 [97.4, 100.0]	59.0	0.5 [0.0, 1.5]	1
	GPT-4o-mini	61.5 [56.5, 66.5]	99.2 [97.4, 100.0]	61.0	0.5 [0.0, 1.5]	1
CORD	Haiku	90.4 [86.8, 93.7]	95.1 [92.6, 97.3]	86.0	4.4 [2.5, 6.6]	16
	GPT-4o-mini	76.9 [72.6, 81.0]	96.8 [94.7, 98.6]	74.4	2.5 [1.1, 4.2]	9
W-2	Haiku	41.3 [38.0, 44.7]	75.3 [70.6, 79.8]	31.1	**10.2** [**8.5**, **12.0**]	194
	GPT-4o-mini	56.1 [54.2, 57.8]	89.2 [86.7, 91.5]	50.0	6.1 [4.8, 7.4]	115
LEDGAR	Haiku	100.0 [100.0, 100.0]	78.1 [75.3, 80.9]	78.1	**21.9** [**19.1**, **24.7**]	164
	GPT-4o-mini	100.0 [100.0, 100.0]	75.1 [71.9, 78.1]	75.1	**24.9** [**21.9**, **28.1**]	187
CUAD	Haiku	100.0 [100.0, 100.0]	74.5 [72.0, 77.2]	74.5	**25.5** [**22.8**, **28.0**]	191
	GPT-4o-mini	100.0 [100.0, 100.0]	66.7 [63.7, 69.6]	66.7	**33.3** [**30.4**, **36.3**]	250
VRDU Reg.	Haiku	77.4 [70.7, 83.7]	99.4 [98.1, 100.0]	77.0	0.4 [0.0, 1.4]	1
	GPT-4o-mini	73.9 [67.6, 80.1]	100.0 [100.0, 100.0]	73.9	0.0 [0.0, 0.0]	0
VRDU Ad-buy	Haiku	77.5 [70.7, 83.5]	100.0 [100.0, 100.0]	77.5	0.0 [0.0, 0.0]	0
	GPT-4o-mini	64.3 [57.8, 70.7]	100.0 [100.0, 100.0]	64.3	0.0 [0.0, 0.0]	0
GST	Haiku	29.2 [24.2, 34.8]	96.9 [94.2, 99.3]	28.3	0.9 [0.2, 1.7]	5
	GPT-4o-mini	30.5 [25.4, 35.9]	98.8 [97.0, 100.0]	30.1	0.4 [0.0, 0.9]	2

4 EXPERIMENTS

4.1 MAIN RESULTS

Table 2 reports Rec_{val}, CBA_{cond}, ACC_{joint}, Δ, and misbinding counts, with 95% paired document-level bootstrap confidence intervals (10,000 resamples; Appendix G). Figure 2 plots Δ. All runs are scored with Algorithm 1, and CORD reflects its LLM-vision track only. Three regimes emerge; Appendix C gives the details.

High misbinding. CUAD has the largest Δ (25.5–33.3%): perfect Rec_{val} masks 191 (Haiku) and 250 (GPT-4o-mini) misbindings. LEDGAR is close behind (Δ 21.9–24.9%).

Misbinding under partial recovery. W-2 (Δ 6.1–10.2%) and CORD (2.5–4.4%) combine incomplete extraction (Rec_{val} 41–90%) with real binding error among the values that are recovered (CBA_{cond} 75–97%). The naive $Rec_{val} - CBA_{cond}$ would be negative for both.

Low-ambiguity baselines. SROIE, both VRDU subsets, GST, and Paystubs show $\Delta \leq 0.9\%$ even where Rec_{val} is low (GST: $Rec_{val} \approx 30\%$, CBA_{cond} 97–99%). CBA, therefore, does not count extraction difficulty as misbinding.

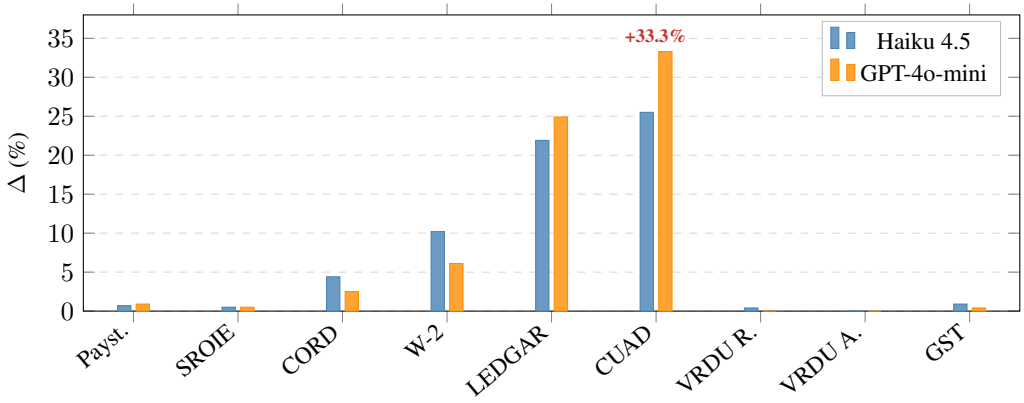


Figure 2: Delta ($\Delta = Rec_{val} - ACC_{joint}$) across domains, computed with the Hungarian-matching scorer (Algorithm 1, `cba_scorer.py`). $\Delta \geq 0$ by construction (Section 3.1), so the y-axis starts at 0. CORD reflects its LLM-vision track only (see Appendix E).

4.2 Where Misbinding Concentrates

Concept density (W-2). Despite its fixed layout, W-2's 44 fields yield 194 (Haiku) and 115 (GPT-4o-mini) misbindings. For Haiku, 89% stay within one ontology family, and 77% stay within the finer categories of Figure 3 (Appendix C). The largest clusters come from the form's two mirrored state/local blocks (e.g., `state_2_income_tax` → `state_1_income_tax`, 17 cases) and from `employee_ssn` ↔ `employer_ein` (24 cases in both directions).

Semantic overlap (legal contracts). CUAD's 15 clause types yield 191 (Haiku) and 250 (GPT-4o-mini) misbindings, or 29.4 per concept for both models combined, against 7.0 for W-2. Near-synonym pairs dominate: "Renewal Term" vs. "Notice Period To Terminate Renewal" (29/24 confusions for Haiku/GPT-4o-mini), "Cap On Liability" vs. "Uncapped Liability" (22/35), and "Agreement Date" vs. "Effective Date" (19/29). In LEDGAR, 94% of "Withholdings" provisions mention "tax", blurring the boundary with "Tax Withholdings".

Low extraction need not mean low binding reliability. A near-zero Δ can come from reliable binding or, in principle, from recovering almost nothing, so Rec_{val} and CBA_{cond} must be read together. SROIE (Rec_{val} 59–62%), Paystubs (63–71%), and GST (≈30%, from 640×640 images) all keep $CBA_{cond} \geq 97\%$. CORD is the partial exception: its CBA_{cond} is 95–97%, and its Δ (2.5–4.4%) is larger than that of any low-ambiguity domain.

4.3 Model Comparison and Open-Weights Generalization

The two proprietary models have complementary strengths. On legal text both reach 100% Rec_{val}, but Haiku binds better (CUAD CBA_{cond} 74.5% vs. 66.7%; LEDGAR 78.1% vs. 75.1%). On W-2 the pattern reverses: GPT-4o-mini extracts more (Rec_{val} 56.1% vs. 41.3%), misbinds less (115 vs. 194), and binds better (CBA_{cond} 89.2% vs. 75.3%). Because Rec_{val} treats the two models as identical on CUAD, CBA_{cond} adds the resolution needed for domain-specific model selection.

To test whether misbinding is specific to proprietary models, we run Qwen2-VL-7B-Instruct (4-bit, run locally) on CUAD, W-2, SROIE, and both VRDU subsets with the same samples and prompts (Table 3; full analysis in Appendix C.6). On CUAD its CBA_{cond} is 42.9%, with 428 misbindings, over 1.7× GPT-4o-mini's count. In the low-ambiguity domains, its CBA_{cond} remains at 97.2–99.4% with $\Delta \leq 2\%$. Its W-2 Rec_{val} is much lower (23.9%), but its W-2 CBA_{cond} (85.5%) exceeds Haiku's, which is consistent with weaker extraction rather than weaker binding on dense forms.

Table 3: Qwen2-VL-7B-Instruct (4-bit) evaluated with the same sampling and prompts as Table 2.

Domain	Rec_{val}	CBA_{cond}	ACC_{joint}	Δ	**Misbindings**
CUAD	100.0%	42.9%	42.9%	57.1%	428
SROIE	72.0%	97.2%	70.0%	2.0%	4
W-2	23.9%	85.5%	20.5%	3.5%	66
VRDU Reg.	77.0%	99.4%	76.5%	0.4%	1
VRDU Ad-buy	62.5%	98.0%	61.3%	1.3%	5

Table 4: Free JSON vs. schema-constrained prompting, same 50-document sample per domain.

Domain	**Model**	**Condition**	Rec_{val}	CBA_{cond}	**Misbindings**
CUAD	Haiku	Free JSON	100.0%	74.5%	191
CUAD	Haiku	Schema-Constrained	88.4%	71.3%	190
CUAD	GPT-4o-mini	Free JSON	100.0%	66.7%	250
CUAD	GPT-4o-mini	Schema-Constrained	84.4%	61.6%	243
W-2	Haiku	Free JSON	43.6%	65.3%	287
W-2	Haiku	Schema-Constrained	41.3%	75.3%	194
W-2	GPT-4o-mini	Free JSON	36.7%	75.6%	170
W-2	GPT-4o-mini	Schema-Constrained	56.1%	89.2%	115
Paystubs	Haiku	Free JSON	71.1%	99.0%	5
Paystubs	Haiku	Schema-Constrained	75.4%	100.0%	0
Paystubs	GPT-4o-mini	Free JSON	62.5%	98.6%	6
Paystubs	GPT-4o-mini	Schema-Constrained	69.5%	95.9%	20

4.4 Prompt Format Ablation

The main runs use fixed concept keys for W-2, SROIE, and GST, and free JSON for CUAD, LEDGAR, and Paystubs (Section 3.5). We run the alternative format on CUAD, W-2, and Paystubs with identical document samples (Table 4; full analysis in Appendix C.7). The effect depends on the domain. On W-2 (44 concepts), fixed keys raise CBA_{cond} by 10.0 (Haiku) and 13.6 (GPT-4o-mini) points and cut misbindings by about a third. On CUAD, fixed keys slightly lower CBA_{cond} and introduce an extraction shortfall (Rec_{val} 84–88%). On Paystubs, Haiku reaches zero misbindings while GPT-4o-mini rises from 6 to 20. Schema-constrained prompting is therefore only a partial, domain-dependent mitigation. It helps when the concept inventory is large, and it can backfire when fixed keys force low-confidence guesses.

4.5 Controlled Density × Semantic-Overlap Factorial Study

The nine domains confound density and overlap: CUAD combines moderate density with extreme overlap, and W-2 combines high density with more moderate overlap. We therefore run a 5×4 factorial study on W-2. It crosses five density levels ($N \in \{4, 8, 16, 32, 44\}$ active concepts) with four overlap conditions. Each condition is an ordering of the 44 concepts, and the first N form the active set:

- *unrelated*: round-robin across families;
- *same-type*: monetary fields first;
- *same-family*: whole families first;
- *near-synonym*: the confusable pairs of Section 4.2 first.

Each of the 20 cells uses a schema-constrained prompt that lists only its N concepts, on the same 50 forms and with both models. Appendix C.8 gives the full design and the error-rate grid (Figure 4).

Table 5 reports the main effects. Averaging out density, overlap moves CBA_{cond} by 25 points in the hypothesized order. Averaging out overlap, density lowers CBA_{cond} by about 16 points from $N = 4$

Table 5: Main effects from the factorial study; each entry averages over the other factor and both models.

Overlap	CBA_{cond}	Δ	**Density**	CBA_{cond}	Δ
Unrelated	87.7%	4.9%	$N = 4$	82.7%	5.4%
Same-type	77.6%	9.5%	$N = 8$	81.0%	6.5%
Same-family	73.0%	9.2%	$N = 16$	73.4%	9.1%
Near-synonym	62.9%	12.9%	$N = 32$	66.8%	13.4%
			$N = 44$	72.5%	11.2%

to $N = 32$. Over the same range, total misbindings rise from 86 to 1,550, an $18\times$ increase for an $8\times$ increase in field count. CBA_{cond} recovers slightly at $N = 44$, because the larger recovered-item denominator dilutes the rate. Overlap matters even at low density: at $N = 4$, near-synonym fields drop Haiku's CBA_{cond} to 45.0%, against 93.3% for unrelated fields, while GPT-4o-mini stays robust (98.2%).

5 Discussion

5.1 When Does Misbinding Matter?

Misbinding becomes consequential when three conditions co-occur. First, several fields must share a data type; with a single dollar amount, date, or name, a value cannot be bound to the wrong slot. Second, downstream systems must consume the structured output without human review: a swapped field in a chat response may be noticed, but the same error in an API response that feeds payroll, contract analysis, or tax software propagates directly into decisions. Third, the value must be plausible in more than one slot, as when "123-45-6789" could be an employee SSN or an employer EIN, or "$52,400" could be Social Security or Medicare wages. Such errors pass type, range, and format checks, so only concept-level verification catches them; this is what CBA measures and what Rec_{val} cannot.

Our results also show that conclusions about an extraction system can change depending on whether extraction and binding are measured jointly or separately. Rec_{val} rates Haiku and GPT-4o-mini as identical on CUAD, whereas CBA_{cond} separates them by 7.8 points, and on W-2 the ranking reverses. The factorial study further shows that density and overlap affect the accuracy of binding for each move across identical documents and models. That overlap alone, even with only four active concepts, is enough to cause serious misbinding. Practitioners should therefore report CBA alongside a joint metric and expect schema-constrained prompting to be most helpful when the concept inventory is large, as our ablation shows.

CBA-soft adds an ontology-relative view of error structure. Its gap over CBA-strict is 4.5–11.0 points on W-2, 7.6–9.8 on LEDGAR, and 9.9–12.9 on CUAD, indicating that a substantial share of misbindings remains within a single semantic family. Because the families reflect author judgment rather than measured business impact, we treat this gap as a diagnostic, not a severity scale (Appendix L.1); Appendix D extends this discussion.

5.2 Limitations

Several limitations qualify these findings. Ground truth for GST was constructed using Claude Sonnet 4.6, so biases shared with the evaluated models could leave some misbindings undetected; the synthetic Paystubs and W-2 domains provide noise-free anchors to mitigate this risk. The ontology families reflect the author's judgment: they were finalized before any model was evaluated, but were not pre-registered externally, so CBA-soft should be read as ontology-relative. In CUAD, inter-annotator disagreement affects 5–10% of labels (Hendrycks et al., 2021), so some apparent misbindings may reflect genuine multi-label ambiguity, and CUAD counts are best read as upper bounds.

Our coverage is also limited. We evaluate two cost-efficient proprietary models and one open-weights model; the latter is tested on five of the nine domains, and settings such as medical records, insurance

forms, and banking statements remain untested. Absolute misbinding rates depend on the prompt format (Section 4.4), so they should not be compared across studies that use different prompts. However, the cross-domain patterns are likely more robust. Finally, API outputs vary slightly even at $T{=}0$ and across routing infrastructure; we therefore treat large misbinding gaps as reliable and near-zero deltas as no meaningful difference. Appendix D.2 gives the full discussion.

6 Conclusion

Value-only evaluation assumes that if the right values appear in the output, extraction is correct. This assumption fails whenever a document contains several fields of the same type, which is the norm for tax forms, contracts, and invoices. We introduced Concept-Binding Accuracy, a metric family that separates value extraction from concept binding, and Hindsight, a benchmark that applies it to nine datasets across five domains with hierarchical concept ontologies under a value-first matching protocol.

Our evaluation shows that Rec_{val} overstates joint key-value accuracy by up to 33.3 percentage points, most sharply on legal contracts, where GPT-4o-mini's perfect value recall hides a misbinding in one of every three extracted clauses. Fixed layouts do not prevent the failure: standardized W-2 forms still produce 194 (Haiku) and 115 (GPT-4o-mini) misbindings when 44 same-type fields compete. Semantic overlap is a stronger per-concept driver than density, and a controlled factorial study confirms both drivers on identical documents and models. The same pattern appears for an open-weights model, while low-ambiguity domains show that CBA does not count extraction difficulty as misbinding.

Document AI systems that process documents with several same-type fields should therefore report CBA alongside Rec_{val} or schema-aware F1, because binding errors pass format checks and propagate silently into payroll, legal, and compliance decisions. We release all ontologies, scoring code, and notebooks to support this practice. Future work includes evaluating larger models, developing mitigation strategies beyond prompt design, and extending Hindsight to new document domains.

AI Usage Statement

In accordance with the ICLR policy on the use of large language models (LLMs), we disclose that LLMs were used in the following ways during this work. (i) *Writing*: LLMs were used to aid and polish writing, including grammar, clarity, and consistency of terminology, and (ii) to draft portions of sections, which the authors then substantially revised. (iii) *Retrieval and discovery*: LLMs assisted in finding and summarizing related work; the authors checked every cited work against the source. (iv) *Research ideation and execution*: LLMs were used to discuss experimental designs and to assist in writing and debugging the experiment code, scoring scripts, and analysis notebooks; all code was reviewed by the authors and validated with unit tests before use. (v) *Mathematical claims*: LLMs were used to help formulate and check the formal definitions and properties of the proposed metrics (e.g., the relationships among Rec_{val}, CBA_{cond}, ACC_{joint}, and Δ in Section 3); the authors verified every derivation. Separately from these uses, LLMs are the *subject* of our evaluation (Claude Haiku 4.5, GPT-4o-mini, and Qwen2-VL-7B-Instruct), and Claude Sonnet 4.6 was used to construct ground truth for the GST Invoices domain, as described in Appendix E and discussed as a limitation in Section 5.2. The authors take full responsibility for all content, including the correctness of text, code, results, and mathematical statements, and for ensuring that no content constitutes plagiarism or scientific misconduct.

Ethics Statement

All authors have read and adhere to the ICLR Code of Ethics. This work involves no human-subject research, crowdsourcing, or collection of new data from individuals, so no IRB review was required. The two domains containing personally identifying field types (Paystubs and W-2 Forms) are fully synthetic: all names, addresses, SSN/EIN values, and monetary amounts are randomly generated and correspond to no real person or organization. All other datasets (SROIE, CORD, LEDGAR, CUAD, VRDU, GST Invoices) are publicly released research datasets used under their stated licenses (Appendix I); where they contain names of real businesses (e.g., receipt merchants or public SEC

and FARA filers), we use this content only as originally released, do not link it to external data, and redistribute only our ontologies, concept annotations, scoring code, and synthetic documents. Proprietary models (Claude Haiku 4.5, GPT-4o-mini) were accessed via commercial APIs under their providers' terms of service, and we note that GST ground truth was constructed using Claude Sonnet 4.6, a potential source of shared-model bias (Section 5.2). We aim to reduce harm: concept misbinding in payroll, tax, and contract pipelines passes format and range checks and can silently propagate into financial, legal, and compliance decisions, and CBA makes this failure measurable. CBA is intended to complement, not replace, human review in high-stakes settings, and a high Hindsight score does not certify a system as safe for deployment, since the benchmark is small, U.S.- and English-centric in its tax and payroll domains, and relies on author-constructed ontologies. We see no direct pathway for misuse of the metric or benchmark. To support research integrity, we report all prompts, compute, and costs (Appendices F and H) and include mixed and negative results alongside positive findings.

Reproducibility Statement

We have taken several steps to make all results in this paper reproducible. All code, concept ontologies, annotations, synthetic documents, and experiment notebooks are available in an anonymized repository at `https://anonymous.4open.science/status/Hindsight-7888` and will be released publicly upon acceptance. The metric definitions (Rec_{val}, CBA_{cond}, ACC_{joint}, CBA-strict/soft, and Δ) are given formally in Section 3, and the value-first Hungarian-matching procedure is specified in Algorithm 1 and implemented in `cba_scorer.py`. Each domain's ontology, including concept families and `confusable_with` annotations, is provided as `data/<domain>/ontology.yaml` (Section 3.4), and dataset sources, sampling, and preprocessing are described in Appendix E, with licenses in Appendix I. The synthetic Paystub and W-2 generation pipelines are included with their generation parameters so that every document can be regenerated. We evaluate pinned model versions (`claude-haiku-4-5-20251001`, `gpt-4o-mini-2024-07-18`, and Qwen2-VL-7B-Instruct at 4-bit via `mlx-vlm`) at temperature T=0 with `random.seed(42)` for all sampling (Section 3.5); full prompts, rate-limiting, and retry settings are listed in Appendix F, and the statistical procedure (10,000-resample paired document-level bootstrap) in Appendix G. Compute, API routing, OCR versions (Tesseract 5.5.3), and per-experiment costs are reported in Appendix H; the full run of all nine datasets reported in Table 2 cost $4.68, so every result can be reproduced cheaply. Because proprietary API outputs may vary slightly across runs, even at T=0, we release the raw model outputs behind every reported number, so all tables and figures can be regenerated exactly from those outputs without re-querying the APIs.

References

Udi Barzelay, Ophir Azulai, Inbar Shapira, Idan Friedman, Foad Abo Dahood, Madison Lee, and Abraham Daniels. Varex: A benchmark for multi-modal structured extraction from documents, 2026. URL `https://arxiv.org/abs/2603.15118`.

Nick Ferguson, Josh Pennington, Narek Beghian, Aravind Mohan, Douwe Kiela, Sheshansh Agrawal, and Thien Hang Nguyen. Extractbench: A benchmark and evaluation methodology for complex structured extraction, 2026. URL `https://arxiv.org/abs/2602.12247`.

Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. Context-dependent fine-grained entity type tagging, 2016. URL `https://arxiv.org/abs/1412.1820`.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. Is information extraction solved by ChatGPT? an analysis of performance, evaluation criteria, robustness and errors, 2023. URL `https://arxiv.org/abs/2305.14450`.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: an expert-annotated NLP dataset for legal contract review. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/6ea9ab1baa0efb9e19094440c317e21b-Abstract-round1.html`.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520, 2019. doi: 10.1109/ICDAR.2019.00244.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pp. 1–6, 2019. doi: 10.1109/ICDARW.2019.10029.

Heng Ji, Ralph Grishman, and Hoa Trang Dang. Overview of the tac 2011 knowledge base population track. In *Proceedings of the Text Analysis Conference (TAC 2011)*, Gaithersburg, Maryland, USA, 2011. National Institute of Standards and Technology (NIST). URL https://tac.nist.gov/publications/2011/papers.html.

Renyu Li, Antonio Jimeno Yepes, Yao You, Kamil Pluciński, Maximilian Operlejn, and Crag Wolfe. Score: A semantic evaluation framework for generative document parsing, 2025. URL https://arxiv.org/abs/2509.19345.

Xiao Ling and Daniel S. Weld. Fine-grained entity recognition. In Jörg Hoffmann and Bart Selman (eds.), *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*, pp. 94–100. AAAI Press, 2012. doi: 10.1609/AAAI.V26I1.8122.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A Dataset for VQA on Document Images . In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208, Los Alamitos, CA, USA, January 2021. IEEE Computer Society. doi: 10.1109/WACV48630.2021.00225.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. Omnidocbench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24838–24848. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02313.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. CORD: A consolidated receipt dataset for post-OCR parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. URL https://openreview.net/forum?id=SJl3z659UH.

Kalyani Roy, Tapas Nayak, and Pawan Goyal. Exploring generative models for joint attribute value extraction from product titles, 2022. URL https://arxiv.org/abs/2208.07130.

Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, November 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2646371. URL https://doi.org/10.1109/TPAMI.2016.2646371.

Anubhav Shrimal, Aryan Jain, Soumyajit Chowdhury, and Promod Yenigalla. PARSE: LLM driven schema optimization for reliable entity extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2749–2763, Suzhou, China, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.emnlp-industry.184/.

Stepán Simsa, Milan Sulc, Michal Uricár, Yash Patel, Ahmed Hamdi, Matej Kocián, Matyás Skalický, Jirí Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. Docile benchmark for document information localization and extraction. In Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi (eds.), *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part II*, Lecture Notes in Computer Science, pp. 147–166. Springer, 2023. doi: 10.1007/978-3-031-41679-8_9.

Tomasz Stanislawek, Filip Gralinski, Anna Wróblewska, Dawid Lipinski, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemyslaw Biecek. Kleister: Key information extraction datasets involving long documents with complex layouts. In Josep Lladós, Daniel Lopresti, and Seiichi Uchida (eds.), *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, Lecture Notes in Computer Science, pp. 564–579. Springer, 2021. doi: 10.1007/978-3-030-86549-8_36.

Stacey Svetlichnaya. Deepform: Understand structured documents at scale, October 2020. URL https://wandb.ai/stacey/deepform_v1/reports/DeepForm-Understand-Structured-Documents-at-Scale--VmlldzoyODQ3Njg. Weights & Biases report. Created October 19, 2020; updated November 21, 2020.

Sönke Tenckhoff, Mario Koddenbrock, and Erik Rodner. Llmstructbench: Benchmarking large language model structured data extraction, 2026. URL https://arxiv.org/abs/2602.14743.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://aclanthology.org/W03-0419/.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pp. 1235–1241. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.155/.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. VRDU: A benchmark for visually-rich document understanding. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye (eds.), *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pp. 5184–5193. ACM, 2023. doi: 10.1145/3580305.3599929.

Writer Information. Tax invoice dataset. https://universe.roboflow.com/writer-information-6sirk/tax-invoice-ud2jo, August 2023. URL https://universe.roboflow.com/writer-information-6sirk/tax-invoice-ud2jo. CC BY 4.0 license. Visited on 2026-04-22.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 1192–1200, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403172.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Fan Yang. OpenTag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2670–2679. ACM, 2018. doi: 10.1145/3219819.3220039.

A Extended Related Work

This appendix gives the full related-work discussion condensed in Section 2.

Most document AI benchmarks and metrics emphasize whether a system can *read* the correct text span, not whether it assigns that span to the correct *semantic concept*. Standard information-extraction evaluation computes precision, recall, and F1 over matched field–value pairs, typically after normalization, and has historically been appropriate when OCR and parsing errors dominated system performance (Tjong Kim Sang & De Meulder, 2003; Shi et al., 2017; Xu et al., 2020). This focus is also reflected in widely used document benchmarks. FUNSD and VRDU evaluate form understanding with standard extraction-style metrics (Jaume et al., 2019; Wang et al., 2023); SROIE measures a small set of well-separated receipt fields using exact-match F1 (Huang et al., 2019); DocVQA frames extraction as question answering, so each target field is scored independently rather than against competing alternatives within a schema (Mathew et al., 2021); and CORD evaluates receipt parsing with structure-aware scoring that conflates structural and semantic assignment errors (Park et al., 2019). CUAD similarly evaluates legal clause retrieval via span overlap and retrieval metrics, which capture whether a relevant clause was found but not whether it was bound to the correct clause category when near-synonymous concepts compete (Hendrycks et al., 2021). Recent benchmarks such as DocILE, OmniDocBench, and SCORE-Bench push evaluation toward large-scale structured extraction and perception-layer fidelity, but they still do not isolate the specific failure mode in which the correct value is extracted and attached to the wrong concept (Simsa et al., 2023; Ouyang et al., 2025; Li et al., 2025).

A related line of work evaluates schema-following and the generation of structured outputs. Extract-Bench, VAREX, and related schema-driven benchmarks stress whether models can produce valid JSON under complex schemas and large field inventories (Ferguson et al., 2026; Barzelay et al., 2026; Tenckhoff et al., 2026). These benchmarks are important for measuring format validity and completeness. However, their scores still aggregate extraction and assignment behavior, making it difficult to determine whether an error arose from failing to find the value or from confusing two semantically adjacent fields.

A closely related body of work targets structured attribute and slot extraction directly. OpenTag formulates attribute-value extraction as sequence tagging over product profiles, but assumes the target attribute is given as input rather than predicted jointly with the value (Zheng et al., 2018). Subsequent generative approaches instead produce (attribute, value) pairs directly via sequence generation, but score them with exact-match accuracy over the joint pair, which, like Rec_{val}-style baselines, does not decompose whether an error originates in value generation or in attribute (concept) assignment (Roy et al., 2022). Large-scale benchmarking of LLM-based information extraction similarly reports aggregate F1 across many IE sub-tasks without isolating extraction failures from assignment failures (Han et al., 2023). Most closely related to our schema-design framing, PARSE shows that the design of the output JSON schema itself, e.g., ambiguous or overlapping field keys, can induce systematic LLM extraction errors, offering a schema-design perspective on the same concept-density and semantic-overlap drivers that we study empirically at the benchmark level (Shrimal et al., 2025).

More generally, adjacent NLP evaluation settings provide only partial precedents. Entity-level F1 in information extraction includes the predicted type as a matching criterion. However, the type systems are typically coarse, and evaluation is not designed to separate extraction from within-document concept assignment errors (Tjong Kim Sang & De Meulder, 2003). Fine-grained entity-type tagging assigns instances to one of dozens or hundreds of fine-grained types drawn from a shared taxonomy, and label-confusion analyses in this setting examine which types are systematically confused (Ling & Weld, 2012; Gillick et al., 2016); CBA asks the analogous question at the level of a single document's schema, whether an extracted value was bound to the correct *document-local* concept, rather than which types are globally confusable across a corpus-wide taxonomy. Entity linking and TAC-KBP slot-filling likewise consider disambiguation and slot semantics, but operate across corpora and knowledge bases rather than over competing fields within a single document extraction output (Ji et al., 2011); CBA is, in effect, a within-document, single-schema analog of this slot-disambiguation problem.

Our work is motivated by the observation that this distinction matters operationally. In modern document pipelines, an extraction error is often detectable through formatting, range, or reconciliation

checks. In contrast, a misassignment between two plausible, same-type fields can pass downstream validation while silently altering the output's meaning. It is especially consequential in domains such as payroll, legal contract review, and invoice processing, where neighboring fields may share the same surface type but carry materially different semantics. Hindsight and CBA are designed to evaluate this missing layer directly: not just whether a model extracted the right value, but whether it bound that value to the right concept.

B Extended Method Details

B.1 Relationship to Schema-Aware Key-Value F1 (Extended)

A standard schema-aware key-value F1 evaluates joint tuples (c_i, v_i) directly: a prediction counts as a true positive only if both the concept key and the value match, and a misbinding is scored as one false negative (the ground-truth tuple was never produced) and one false positive (an incorrect tuple was produced instead). Standard key-value F1 therefore *does* penalize misbinding; the value-only baseline Rec_{val} (Section 1) does not. What Rec_{val} and CBA jointly provide, that schema-aware F1's single scalar does not, is a *decomposition* of joint extraction error into two independent components: (a) value-reading or OCR failure $(1 - Rec_{val})$ and (b) concept-assignment failure among recovered values $(1 - CBA_{cond})$. A schema-aware F1 score of, say, 66.7% on CUAD is consistent with many different underlying failure profiles; CBA's decomposition ($Rec_{val} = 100\%$, $CBA_{cond} = 66.7\%$) identifies that the failure profile here is *entirely* concept assignment, with zero contribution from value extraction, a distinction with different practical implications, better prompting versus better OCR, for example, that the joint metric alone cannot support. Table 6 compares Rec_{val}, key-value tuple F1, schema-aware structural F1, hierarchical tree-distance metrics, and CBA_{cond} along the properties relevant to this decomposition.

Table 6: Structured comparison of value-extraction and concept-assignment evaluation metrics. "Isolates" means the metric can be computed, or read off, independently of the other axis; "diagnostic decomposition" means the metric family as a whole supports attributing a joint error to a specific cause.

Metric	Isolates value extraction?	Isolates concept assignment?	Requires schema pre-allocation?	Evaluates structural validity?	Diagnostic decomposition?
Value-Only Recall (Rec_{val})	Yes	No	No	No	No (extraction only)
Key-Value Tuple F1 ($F1_{tuple}$)	No	No	No	No	No (joint scalar only)
Schema-Aware Structural F1 ($F1_{schema}$)	No	No	Yes	Partial	No (joint scalar only)
Hierarchical Tree Distance	No	Partial	Yes	Yes	No (single cost scalar)
Concept-Binding Accuracy (CBA_{cond})	No (conditional on recovery)	Yes	No	No	Yes, jointly with Rec_{val}

B.2 Value-First Protocol: Step-by-Step Rules

Informally, Algorithm 1 (Section 3.3) applies the following rules:

i) For each (c_i, v_i), search predictions for a matching value v'_j (norm).

ii) If found under $c'_j = c_i$: correct extraction and binding (counts toward Rec_{val} and CBA_{cond}'s numerator).

iii) If found under $c'_j \neq c_i$: correct extraction, misbinding (counts toward Rec_{val}'s numerator only, i.e., a CBA_{cond} miss).

iv) If multiple candidate matches exist for a value, the Hungarian assignment prefers the exact concept match (lowest cost) over any other candidate.

v) If no value match exists (complete miss), the item does not count toward Rec_{val} and is excluded from CBA_{cond}'s denominator.

B.3 Ontology Construction and Design Space

The authors constructed each family grouping following three criteria: (a) *shared surface type* (e.g., all dollar-amount wage fields), (b) *co-occurrence in the same document region or clause*, and (c) *plausible confusability*, whether an annotator or model could reasonably assign

a value to either concept given only its type and local context, documented per concept as the `confusable_with` field in each domain's ontology file (`data/<domain>/ontology.yaml`; see `docs/adding_domains.md`). We flag explicitly that this process involves researcher judgment: two reasonable domain experts could draw family boundaries differently (e.g., whether `state_wages` belongs with federal wage fields or with the state-specific block), and the resulting CBA-soft partial-credit gap (Section L.1) is a function of these choices, not solely of model behavior. This is precisely why we present CBA-soft as an ontology-relative diagnostic rather than a calibrated severity score. For benchmark deployments where the ontology influences reported results, particularly comparative claims across models, we recommend pre-registering the ontology (family groupings and confusability annotations) before running any model evaluation, so that family boundaries cannot be adjusted, even unintentionally, in response to observed model errors. Hindsight's own ontologies were finalized before any model evaluation was run, but we did not pre-register them externally, and we flag this as a limitation (Section 5.2).

Table 7 places each domain in the concept-density $\times$ semantic-overlap design space, and Table 8 summarizes each domain's key binding challenge; full domain descriptions are in Appendix E.

Table 7: Density $\times$ Overlap design space.

Density / Overlap	**Low Semantic Overlap**	**High Semantic Overlap**
Low density	SROIE (4)	—
Moderate density	Paystubs (14), CORD (13), VRDU (6–9), GST (13)	LEDGAR (15), CUAD (15)
High density	—	W-2 (44)

Table 8: High-level domain characteristics.

Domain (source)	**Key binding challenge**
Paystubs (syn.)	Distinct labels, low ambiguity baseline; 200 docs.
SROIE (Huang et al., 2019)	4 distinct-type fields; binding nearly trivial.
CORD (Park et al., 2019)	Indon. receipts; extraction noise dominates.
W-2 (syn.)	44 fixed-layout fields; highest concept density.
LEDGAR (Tuggener et al., 2020)	Employment clauses; near-synonyms (Withholdings vs Tax Withholdings).
CUAD (Hendrycks et al., 2021)	Commercial contracts; extreme semantic overlap (Cap vs Uncapped).
VRDU (Wang et al., 2023)	Government forms; clearly separated fields.
GST Invoices (Writer Information, 2023)	Tax trio (CGST/SGST/IGST); low-resolution images.

C Extended Experimental Results

This appendix gives the full analyses summarized in Section 4.

C.1 Cross-Domain Observations

Key observations from Table 2:

- **High misbinding domains:** CUAD shows the largest Δ (33.3% for GPT-4o-mini), with perfect Rec_{val} masking 250 misbindings. LEDGAR shows a comparably large Δ (21.9–24.9%). Because $Rec_{val} = 100\%$ for both, $ACC_{joint} = CBA_{cond}$ and $\Delta = 1 - CBA_{cond}$ here.
- **Moderate misbinding under partial recovery:** W-2 ($\Delta = 6.1$–10.2%) and CORD ($\Delta = 2.5$–4.4%) combine incomplete extraction (Rec_{val} 41–90%) with real, non-trivial binding error among the values that *are* recovered (CBA_{cond} 75–97%). The naive difference

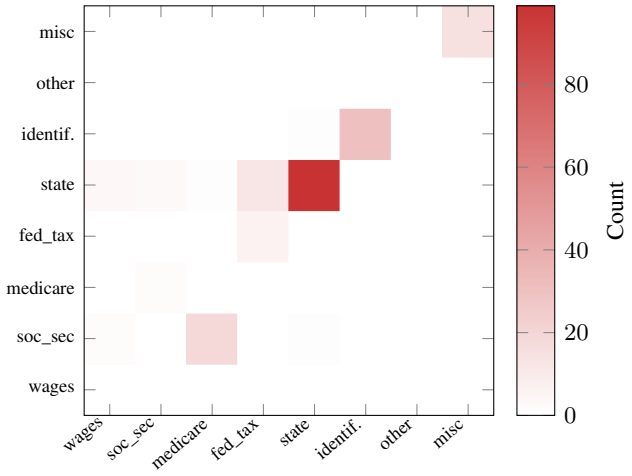


Figure 3: W-2 concept confusion matrix (Claude Haiku 4.5). Rows are the ground-truth field's category, columns the category it was bound to; cell (i, j) counts misbindings from category i to category j. 77% of Haiku's 194 misbindings fall on the diagonal (within-category); the `state` category (mirrored state/local reporting blocks) accounts for the single largest cluster (99 within-category, plus 20 state values bound to `fed_tax` or `wages`/`soc_sec`/`medicare`).

$Rec_{val} - CBA_{cond}$ would be negative for both domains; Δ instead shows a genuine, if smaller, misbinding signal in both.

- **Low-ambiguity baselines:** SROIE, VRDU Registration, VRDU Ad-buy, GST, and Paystubs all show $\Delta \leq 0.9\%$, confirming that CBA does not penalize when misbinding is genuinely absent, *even when Rec_{val} itself is low* (GST: $Rec_{val} \approx 30\%$, $CBA_{cond} \approx$ 97–99%). Because CBA_{cond} reports binding reliability directly, a domain's extraction difficulty (low Rec_{val}) need not be disentangled from its binding difficulty by inspecting the sign of a difference, and Δ stays near zero whenever binding reliability is high, regardless of Rec_{val}.

C.2 Concept Density: The W-2 Effect

W-2 forms contain 44 fields in a fixed layout yet produce 194 misbindings for Claude Haiku 4.5 and 115 for GPT-4o-mini (309 combined, reported separately per model to avoid inflating counts by summing errors on identical documents), indicating that high concept density alone can create substantial binding difficulty. Of Haiku's 194 misbindings, 89% occur within the same ontology family (Section 3.4); at the finer visualization granularity of Figure 3, which splits the `compensation` family into `wages`/`soc_sec`/`medicare` sub-categories, 77% remain within-category. The two most frequent individual pairs are `employee_ssn` $\leftrightarrow$ `employer_ein` (24 bidirectional occurrences) and `social_security_tips` $\rightarrow$ `medicare_wages` (16 occurrences), together with a cluster of state/local field swaps (`state_2_income_tax` $\rightarrow$ `state_1_income_tax`, 17 occurrences; `state_2` $\rightarrow$ `state_1` and `state_2_wages` $\rightarrow$ `state_1_wages`, 14 each) reflecting the form's two mirrored state-reporting blocks. Figure 3 shows this concentration: The `state` category alone accounts for 99 within-category misbindings plus a further 20 cross-category misbindings in which ground-truth state/local values were bound to `fed_tax` (12) or `wages`/`soc_sec`/`medicare` (8 combined) fields; the within-category cluster is consistent with `local_*_income_tax` fields being bound to the near-identical-looking `state_*_income_tax` fields. The `identif` category (30 within-category misbindings, dominated by the SSN/EIN pair above) is the second largest cluster.

C.3 Semantic Overlap: Legal Contract Effect

Legal domains show the highest misbinding rate per concept. CUAD's 15 clause types produce 191 misbindings for Claude Haiku 4.5 and 250 for GPT-4o-mini (reported per model to avoid inflating counts by summing errors on identical documents; the two-model total of 441 corresponds

to 29.4 per concept), driven by near-synonym pairs such as "Renewal Term" vs. "Notice Period To Terminate Renewal" (29 bidirectional confusions for Haiku, 24 for GPT-4o-mini), "Cap On Liability" vs. "Uncapped Liability" (22 for Haiku, 35 for GPT-4o-mini), and "Agreement Date" vs. "Effective Date" (19 for Haiku, 29 for GPT-4o-mini). In LEDGAR, "Withholdings" vs. "Tax Withholdings" accounts for a substantial share of errors; 94% of "Withholdings" provisions mention "tax," making the distinction especially subtle.

C.4 Baseline Validation: Low Extraction Need Not Mean Low Binding Reliability

Δ is not a signed indicator that separates "extraction-dominated" from "binding-dominated" domains (Section 3.1): since $\Delta \geq 0$ always, a near-zero Δ can arise either because a domain is genuinely low-ambiguity (high CBA_{cond} regardless of Rec_{val}) or, in principle, because almost nothing was recovered at all. Disentangling these two cases requires reading Rec_{val} and CBA_{cond} together, which is exactly what the three-metric framework is designed to support. Five of our nine domains show this "good binding despite incomplete extraction" pattern clearly: SROIE ($Rec_{val} = 59$–62%, $CBA_{cond} \approx 99\%$), VRDU registration and ad-buy forms (at most one misbinding per model out of up to 50 documents), Paystubs ($Rec_{val} = 63$–71%, $CBA_{cond} \approx 99\%$), and GST. On GST, low-resolution 640×640 images reduce Rec_{val} to about 30%, while CBA_{cond} remains 97–99%, showing that binding is highly reliable among the values that *are* recovered even though most values are never recovered at all; the domain's near-zero Δ (0.4–0.9%) reflects this directly, since $\Delta = Rec_{val} \times (1 - CBA_{cond})$ is small whenever CBA_{cond} is close to 1, independent of how low Rec_{val} itself is. CORD is the partial exception: its CBA_{cond} (95–97%) is still high in absolute terms, but its Δ (2.5–4.4%) is larger than that of any of these five domains, indicating a small but real amount of genuine misbinding alongside CORD's well-documented OCR/extraction noise (Appendix E).

C.5 Model Comparison

The two models show complementary strengths, suggesting that extraction and concept binding are partially independent capabilities. On legal text, both achieve perfect Rec_{val}, but Haiku is the stronger binder: GPT-4o-mini trails by 7.8 percentage points on CUAD (CBA_{cond} 66.7% vs. 74.5%) and by 3.0 points on LEDGAR (75.1% vs. 78.1%). On W-2 forms, the pattern reverses: GPT-4o-mini achieves higher extraction accuracy (Rec_{val} 56.1% vs. 41.3%), fewer misbindings (115 vs. 194), and a 13.9-point higher CBA_{cond} (89.2% vs. 75.3%). On GST, both models struggle with extraction ($Rec_{val} \approx 29$–31%), while binding remains comparatively strong (CBA_{cond} 97–99%). Because Rec_{val} alone treats models as nearly identical on CUAD, CBA_{cond} provides the additional resolution needed for domain-specific model selection.

C.6 Open-Weights Model Generalization

Results are in Table 3 (Section 4.3). Both Haiku 4.5 and GPT-4o-mini are proprietary, cost-efficient production models. To test whether concept misbinding is specific to this model class or a more general property of current vision-language models, we additionally evaluate **Qwen2-VL-7B-Instruct** (4-bit quantized, run locally with zero API cost) on the highest-ambiguity domain (CUAD) and three low-ambiguity domains (SROIE, VRDU Registration, VRDU Ad-buy), plus W-2, using the identical sampling and prompts as the corresponding proprietary-model runs.

Table 3 shows that misbinding is not a proprietary-model-specific artifact, and that the concept-density/semantic-overlap drivers identified in Section 5.1 appear to generalize: on CUAD, Qwen2-VL-7B's CBA_{cond} (42.9%) is more than 20 points below either proprietary model (Haiku 74.5%, GPT-4o-mini 66.7%), with 428 misbindings, over $1.7\times$ GPT-4o-mini's count on the same 50 bundles. On the three low-ambiguity domains (SROIE, VRDU Registration, VRDU Ad-buy), by contrast, Qwen2-VL-7B's CBA_{cond} (97.2–99.4%) is comparable to both proprietary models, and its Δ stays at or below 2%, matching the proprietary models' near-zero misbinding on these domains. Its Rec_{val} is lower than both proprietary models on W-2 (23.9% vs. 41.3%/56.1%) and VRDU Ad-buy (62.5% vs. 77.5%/64.3%), but not uniformly so: it is higher than both on SROIE (72.0% vs. 59.5%/61.5%). The sharp W-2 gap is consistent with a smaller, quantized model being comparatively weaker at raw extraction on dense forms, rather than at concept binding.

C.7 Prompt Format Ablation

Results are in Table 4 (Section 4.4). Our main results use each domain's prompt design, which differs in format across domains: CUAD, LEDGAR, and Paystubs request a free JSON array of `{label, value, concept}` triples (the model both discovers which fields are present and generates the concept name for each), while W-2, SROIE, and GST request a JSON object with pre-allocated, fixed concept keys (the model fills in values for a given schema rather than naming concepts itself). To test whether this design choice itself drives misbinding, independent of domain, we run the alternative condition for three domains, CUAD, W-2, and Paystubs, holding the document sample fixed (identical `random.seed(42)` sampling) and varying only the prompt format.

Table 4 shows the effect is domain-dependent rather than uniform. On W-2, the domain with the most canonical concepts (44), schema-constraining clearly reduces misbinding for both models (CBA_{cond} +10.0 points for Haiku, +13.6 for GPT-4o-mini; misbindings roughly a third lower), consistent with free concept generation amplifying errors when there is a large concept inventory to choose from. On CUAD, schema-constraining does not help and mildly hurts CBA_{cond} for both models, while introducing a new extraction shortfall that free JSON did not have (Rec_{val} drops from 100% to 84–88%, since the model can now fail to assign any clause number to a category, rather than being forced to classify every clause it is given). On Paystubs, the effect is mixed and model-specific: Haiku improves to zero misbindings under the schema constraint, but GPT-4o-mini gets worse (6→20 misbindings), plausibly because being forced to answer all 14 fixed keys pushes it to guess on fields it would otherwise have omitted. We conclude that schema-constrained prompting is not a general mitigation for concept misbinding; it helps specifically when the concept inventory is large enough that free-form concept naming itself becomes error-prone, and can backfire by forcing low-confidence guesses that free-form generation would have skipped.

C.8 Factorial Study: Full Design and Error-Rate Grid

The cross-domain comparison in Section 4.1 attributes misbinding to two drivers, concept density and semantic overlap. However, the nine domains confound them: CUAD combines moderate density with extreme overlap, W-2 combines high density with more moderate overlap, and no domain isolates one factor while holding the other fixed. To test both drivers as independent causal effects on a single, controlled domain, we construct a 5×4 factorial design on W-2: five density levels ($N \in \{4, 8, 16, 32, 44\}$ active concepts) crossed with four semantic-overlap conditions, each defined as an ordering of the 44 W-2 concepts from which the first N are taken as the active set for a given cell:

- **Unrelated**: round-robin across all 8 ontology families, maximizing family spread at low N (at $N = 8$, exactly one concept per family).
- **Same-type**: all 22 monetary ($) fields first, spanning multiple families but sharing surface type, then the 22 non-monetary fields.
- **Same-family**: whole ontology families in descending size order, so `state_local` (the largest, 14 concepts) fills $N \leq 14$ entirely.
- **Near-synonym**: the specific highly-confusable pairs identified in Section 4.2 ordered first, SSN/EIN, the seven mirrored `state_1_*`/`state_2_*` pairs, `local_*_income_tax`, and the `box_12` code/value slots.

Each of the 20 cells uses a schema-constrained extraction prompt listing only that cell's N active concepts, run on the same 50 W-2 forms, and both models are used throughout. Figure 4 shows the resulting error rate ($100 - CBA_{cond}$, averaged across both models) across the full grid: error concentrates toward the top-right (high density, high overlap), with the near-synonym row darker than the other three rows at every density level except $N = 8$ (where same-family is higher).

Table 5 (Section 4.5) shows the two main effects, each averaged over the other factor and both models.

Both drivers hold up as independent causal effects on identical documents and identical models. Averaging out density, overlap alone moves CBA_{cond} by 25 points in the hypothesized order (unrelated > same-type > same-family > near-synonym). Averaging out overlap, density alone moves CBA_{cond} down by roughly 10–16 points from $N = 4$ to $N = 32$, with total misbindings rising from 86 to 1,550 (both models, all four overlap conditions combined), a factor of $18\times$ for an $8\times$ increase in field count. CBA_{cond} ticks back up slightly at $N = 44$ relative to $N = 32$

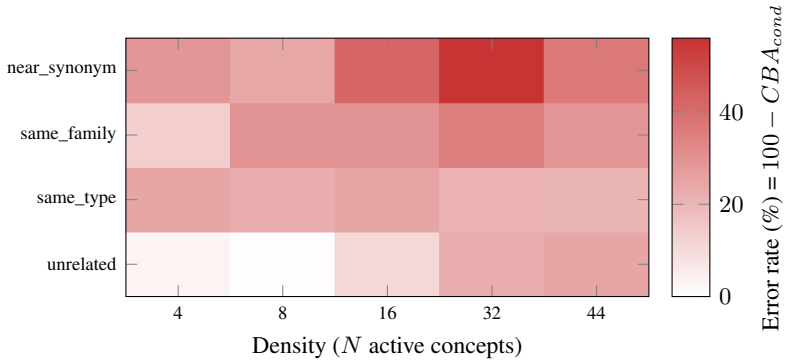


Figure 4: Error rate ($100 - CBA_{cond}$, averaged across both models) across the density × semantic-overlap factorial grid on W-2. Darker cells indicate more misbinding.

for three of the four overlap conditions despite comparable or higher raw misbinding counts: at $N = 44$ the much larger recovered-item denominator dilutes the rate even as absolute errors climb, a distinction between a rate (CBA_{cond}) and a raw count that the two metrics are designed to make visible separately (Section 3.1).

Semantic overlap does not need high density to bite: at $N = 4$, the near-synonym condition (SSN/EIN plus the primary/secondary state code pair) already drops Haiku's CBA_{cond} to 45.0%, far below the unrelated condition's 93.3% at the same density. This is a genuinely independent confirmation that the cross-domain comparison alone could not establish, since no domain in Table 2 isolates high overlap at low density. We also observe a sharp model-specific divergence in this cell: GPT-4o-mini remains robust on the identical four near-synonym fields ($CBA_{cond} = 98.2\%$) where Haiku fails badly, a difference not visible in either model's aggregate performance elsewhere in the paper.

D Extended Discussion

D.1 When Does Misbinding Matter? (Extended)

Our results indicate that concept misbinding becomes a meaningful failure mode when three conditions co-occur. First, *multiple fields must share a data type*: with only one dollar amount, one date, or one name, misbinding is structurally impossible. Risk arises when documents contain several plausible values of the same type, such as multiple wage amounts on a W-2, repeated tax components on a GST invoice, or near-synonym legal clauses. Second, *downstream systems must consume structured outputs without human review*. A misbinding in a chat response may be noticed, but the same error in an API response feeding payroll, contract analysis, or tax software can propagate directly into decisions. Third, *values must be plausible in multiple slots*: for example, "123-45-6789" could denote an employee SSN or employer EIN, and "$52,400" could refer to Social Security wages or Medicare wages. Such values pass type, range, and format checks regardless of assignment, so only concept-level verification catches the error, precisely what CBA measures and Rec_{val} does not. More broadly, these results show that evaluative conclusions about structured extraction systems can change materially depending on whether value extraction and concept assignment are measured jointly or separately. The controlled factorial study (Section 4.5) confirms that the first condition, concept density and semantic overlap, holds as an independent causal effect rather than only a cross-domain correlation: both drivers move CBA_{cond} substantially on identical documents and models, and semantic overlap alone (without high density) is already sufficient to induce serious misbinding.

D.2 Limitations (Extended)

Ground truth. For GST invoices, ground truth was constructed using Claude Sonnet 4.6, so shared model biases could leave some misbindings undetected. Synthetic domains (Paystubs and W-2) provide noise-free anchors.

Ontology construction. Concept families were grouped by the authors (Section 3.4) rather than pre-registered externally before evaluation began;. In contrast, the ontologies were finalized before any model was run, the family boundaries still reflect researcher judgment, and CBA-soft's within/cross-family split should be read as ontology-relative rather than as an externally validated severity scale.

Model coverage. Our main comparison evaluates two cost-efficient proprietary models. We additionally evaluate one open-weights model (Qwen2-VL-7B-Instruct, Section 4.3) on five of nine domains, finding that misbinding generalizes beyond the proprietary-model class and, if anything, is more pronounced on the highest-ambiguity domain tested. Extending the study to flagship models (e.g., GPT-4o, Claude Opus, Gemini 2.5 Pro), larger open-weights models, and the remaining four domains would further strengthen the generalizability of the findings. Based on the pattern observed so far, we expect the qualitative drivers: concept density and semantic overlap, to persist.

Domain coverage. Although nine datasets across five domains provide breadth, important settings such as medical records, insurance forms, and banking statements remain untested. The benchmark is extensible by design, requiring only an ontology and annotated ground truth for new domains.

Prompt sensitivity. Absolute misbinding rates depend on the prompting format. We directly test this (Section 4.4) by running the alternative format (free JSON vs. schema-constrained) on CUAD, W-2, and Paystubs: the effect is real but domain-dependent, helping substantially on the domain with the largest concept inventory (W-2) while being neutral-to-negative on CUAD and mixed on Paystubs. Absolute misbinding rates should therefore not be compared across studies that use different prompting formats, but the relative cross-domain patterns are likely more robust than the absolute rates.

Multi-label ambiguity. In CUAD, some apparent misbindings may reflect legitimate annotation ambiguity; inter-annotator disagreement affects 5–10% of labels (Hendrycks et al., 2021). CUAD misbinding counts should therefore be interpreted as upper bounds.

API non-determinism: Even at $T{=}0$, API outputs can vary slightly between runs. Comparing the reported results (Table 2) against the Rec_{val} and misbinding counts of an earlier run of the same experiments, the largest domains (CUAD, LEDGAR, VRDU) moved by a few points or less, while W-2's per-model numbers shifted somewhat more; we attribute this partly to ordinary run-to-run variance and partly to routing the reported calls through a gateway proxying to AWS Bedrock rather than the direct Anthropic API used in the earlier run (Appendix H), which we cannot fully disentangle without a same-infrastructure rerun. We therefore treat large misbinding gaps as reliable, but interpret near-zero deltas as "no meaningful difference" rather than as directional evidence.

E Detailed Domain Descriptions

Paystubs (synthetic, 200 generated documents, 50 evaluated per model, 14 concepts, 5 families). We generate paystubs via a custom pipeline with controlled field placement and realistic value distributions across 15 layout variants, and evaluate them using a two-stage Tesseract-LLM pipeline. The 14 concepts span compensation (gross pay, net pay, overtime), deductions (federal tax, state tax, Social Security, Medicare), and identifiers (employee name, SSN, pay period). Field labels are distinct and unambiguous: "Gross Pay" never appears near a field that could be confused with it. Therefore, this domain serves as the low-ambiguity baseline where CBA-strict (i.e., CBA_{cond}) should remain high even though Rec_{val} is depressed by OCR failures: concept-binding among recovered values should stay strong. This is confirmed by the results in Table 2: Rec_{val} is 63–71% (OCR misses roughly a third of fields), while CBA_{cond} remains 98–99%, giving a near-zero Δ (0.7–0.9%), the expected signature of a domain where extraction, not binding, is the limiting factor.

SROIE (Huang et al., 2019) (ICDAR 2019, 50 receipts, 4 concepts, 3 families). English-language scanned receipts with four ground-truth fields: company name, date, address, and total amount. Each field has a distinct data type (string, date, address, number), making concept binding nearly trivial. We selected this as the minimal baseline, the smallest concept set in our benchmark, confirming that CBA converges to Rec_{val} when binding ambiguity is absent.

CORD (Park et al., 2019) (CORD v2, 100 receipts, 93 qualifying, 13 concepts, 3 families). Indonesian-language receipts with 13 evaluation concepts, including confusable amount fields

(`subtotal_price` vs. `total_price` vs. `tax_price`) and count fields (`menuqty_cnt` vs. `itemsubtotal_cnt`). We evaluate two extraction tracks: Tesseract OCR followed by LLM concept mapping, and direct LLM vision extraction. This domain tests concept binding under noisy extraction, where Indonesian OCR introduces character errors that compete with binding errors. We report results for the LLM-vision track only (Section 4.1); under it, Rec_{val} is 77–90% and CBA_{cond} 95–97%, giving a larger Δ (2.5–4.4%) than any of the five low-ambiguity domains, evidence of a small but genuine binding error alongside the extraction noise this domain was selected to test.

W-2 Tax Forms (synthetic, 50 forms, 44 concepts, 8 families). IRS W-2 wage and tax statements have the highest concept density in our benchmark: 44 fields including 7 wage amounts, 6 tax withholding amounts, 2 Social Security numbers (employee SSN and employer EIN in different formats), and 2 state reporting blocks with mirrored field structures. Despite the form's fixed layout with clearly labeled boxes, a setting where one might expect binding to be trivial, the sheer number of fields of the same type creates systematic confusion. This domain tests the concept density hypothesis: does misbinding scale with the number of competing same-type fields, even when layout provides explicit labels?

LEDGAR (Tuggener et al., 2020) (LexGLUE/SEC filings, 50 bundles, 15 concepts, 6 families). Employment law contract provisions from SEC EDGAR filings, classified into 15 categories: Terminations, Termination For Cause, Survival, Confidentiality, Non-Compete, Non-Solicitation, Indemnity, Indemnifications, Withholdings, Tax Withholdings, Death, Disability, Positions, Employment, and Governing Law. The text-only format isolates concept binding from extraction noise. Several concept pairs have high semantic overlap: "Withholdings" and "Tax Withholdings" differ by a single qualifier, and in practice 94% of "Withholdings" provisions mention the word "tax." Similarly, "Death" and "Disability" co-occur in termination clauses. This domain tests the semantic overlap hypothesis.

CUAD (Hendrycks et al., 2021) (CUADv1, 50 bundles, 15 concepts, 5 families). Commercial law contracts are used in mergers and acquisitions due diligence, with 15 clause types. CUAD represents the strongest misbinding challenge in our benchmark because it combines moderate concept density with extreme semantic overlap. Key confusion pairs include: "Cap On Liability" vs. "Uncapped Liability" (one limits damages, the other removes limits, the distinction hinges on a single modifier in dense legal prose); "Agreement Date" vs. "Effective Date" (two dates that may or may not coincide); and "Renewal Term" vs. "Notice Period To Terminate Renewal" (one describes the extension period, the other describes the cancellation window). A single contract paragraph may trigger multiple clause types simultaneously, creating genuine multi-label ambiguity that even expert annotators find challenging (Hendrycks et al., 2021).

VRDU (Wang et al., 2023) (Google Research, two sub-domains). VRDU contains registration forms (1,915 documents, 6 concepts, 4 families) from Foreign Agent Registration Act filings, and ad-buy forms (641 documents, 9 concepts, 4 families) from political advertising disclosure filings. Both sub-domains feature well-separated, clearly-labeled fields with minimal semantic overlap. Selected as a structural validation: government forms with moderate concept counts but clear field separation should produce near-zero misbinding, confirming that CBA does not penalize where misbinding is genuinely absent.

GST Invoices (Writer Information, 2023) (Roboflow Tax Invoice dataset, 773 images, 13 concepts, 4 families). Indian Goods and Services Tax invoices with a targeted misbinding challenge: the tax components CGST (Central GST), SGST (State GST), and IGST (Integrated GST). All three are tax fields that frequently carry identical values on the same invoice. Ground truth was constructed using Claude Sonnet 4.6 with guided extraction (the model was told which fields to look for, based on COCO bounding-box annotations), a deliberately different task from the unguided extraction used for evaluation. Images are 640×640 (resized from originals), making OCR unreliable and requiring vision-based extraction.

F Full Extraction Protocol

Prompt design. CUAD, LEDGAR, and Paystubs use a prompt requesting a JSON array of `[{label, value, concept}]` objects, which forces the model to both discover which fields are present

and assign each to a concept; W-2, SROIE, and GST use a JSON object with pre-allocated, fixed concept keys, where misbinding appears as a value placed under the wrong key (Section 4.4). For image-based domains (receipts, W-2, VRDU, GST), base64-encoded images are sent via vision API endpoints. For text-based domains (LEDGAR, CUAD), the document text is passed directly; for Paystubs, Tesseract OCR text is passed to the model (Appendix E). The system prompt provides the list of canonical concepts and brief definitions, but does not indicate which concepts are present in the specific document.

Model details.

- **Claude Haiku 4.5:** `claude-haiku-4-5-20251001`. $T = 0$.
- **GPT-4o-mini:** `gpt-4o-mini-2024-07-18`, OpenAI's cost-efficient production model. $T = 0$.

Reproducibility. All experiments use `random.seed(42)` for sampling, and every experiment is implemented as a self-contained Jupyter notebook with data loading, API calls, scoring, and visualization. Rate limiting: 1.5 s base delay between Anthropic calls, 5 s between OpenAI calls, with exponential backoff on rate-limit errors ($10 \rightarrow 20 \rightarrow 40 \rightarrow 80$ seconds). All notebooks, ontologies, and scoring code are released at `https://anonymous.4open.science/r/hindsight-anonymous-0874/`.

G Statistical Significance

All experiments use temperature-zero (T=0) inference, so reported intervals capture cross-document variability rather than run-to-run noise (residual API non-determinism is discussed in Section 5.2). We report 95% paired document-level bootstrap confidence intervals (10,000 resamples, percentile method, `random.seed(42)`, `scripts/bootstrap_ci.py`) for Rec_{val}, CBA_{cond}, and Δ, shown in brackets in Table 2. With 50 documents per domain (CORD: 93; GST: 100), Rec_{val} 95% CI half-widths are typically $\pm$3–7 percentage points (narrower for LEDGAR/CUAD, where $Rec_{val} = 100\%$ exactly for both models). The headline Δ CIs for CUAD (25.5–33.3%, CI lower bounds 22.8/30.4), LEDGAR (21.9–24.9%, CI lower bounds 19.1/21.9), and W-2 (6.1–10.2%, CI lower bounds 4.8/8.5) all exclude zero, supporting a systematic (not noise-level) misbinding signal on these three domains. The near-zero-Δ domains (SROIE, VRDU Registration, VRDU Ad-buy, GST, Paystubs) mostly do not exclude zero at the 95% level, consistent with genuinely minimal misbinding rather than a systematic effect too small to distinguish from sampling noise. We have provided the implementation in the supplementary notebooks and in `SRC/scripts/` of the released repository.

H Compute Resources

We evaluate two models per domain, Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) and GPT-4o-mini (`gpt-4o-mini-2024-07-18`), and provide a detailed per-domain breakdown of API calls, extraction modality, and wall-clock runtime in Table 9. The total API cost across all reported experiments is approximately $30–55, dominated by vision-based calls for the SROIE, CORD, W-2, VRDU, and GST domains, and no GPU compute costs were incurred. To ensure stable API-based execution, all notebooks implement exponential backoff retry logic with up to 5 attempts and between-request delays of 1-5 seconds, depending on the provider and modality; OpenAI vision calls use the longer 5-second delay due to higher tokens-per-minute consumption.

Compute for the reported runs (Table 2). The results reported in Table 2 and throughout Section 4 were produced end-to-end with the Hungarian-matching implementation (Algorithm 1, `cba_scorer.py`), using each domain's sampling (`random.seed(42)`) and routing model calls through a gateway proxying to AWS Bedrock (Claude Haiku 4.5, Claude Sonnet 4.6 for GST's guided ground-truth construction) and OpenAI (GPT-4o-mini) directly, rather than the direct-API setup described above, which was used for an earlier run of the same experiments. Total cost across all nine datasets was $4.68, substantially below the $30–55 of the direct-API run, reflecting the gateway's routing and pricing rather than a change in the amount of inference performed. The open-weights model evaluation (Section 4, Qwen2-VL-7B-Instruct, 4-bit) ran entirely locally via `mlx-vlm` at zero additional cost. The Paystubs run used a local Tesseract 5.5.3 installation (English-only); CORD's

Table 9: Summary of evaluation datasets, API usage, extraction modality, and wall-clock time per domain. Wall-clock times for VRDU Reg., VRDU Ad-buy, and GST Invoices are estimated from the average per-API-call rate for similar vision-based domains in this benchmark (∼8.6 sec/call); other times are extracted directly from Colab cell execution metadata.

Domain	Docs per Model	Total API Calls	Extraction Approach	Wall Clock Time
Paystubs	50	100	Tesseract OCR + LLM Text	3 mins
Receipts (SROIE)	50	100	LLM vision	13 mins
Receipts (CORD)	93	372	Tesseract+LLM + LLM Vision	45 mins
W-2 Tax Forms	50	100	LLM Vision	20 mins
Employment Law (LEDGAR)	50 bundles	100	LLM-Text only	12 mins
Commercial Law (CUAD)	50 bundles	100	LLM-Text only	12 mins
VRDU Reg.	50	100	LLM Vision	∼14 mins
VRDU Ad-buy	50	100	LLM Vision	∼14 mins
GST Invoices	100	200	LLM Vision	∼29 mins

Tesseract-OCR track was not rerun due to the lack of an Indonesian language pack, so CORD's reported numbers reflect only its LLM-vision track. The prompt-format ablation (Section 4.4, 3 domains $\times$ 1 new condition each) cost an additional $1.1, and the density$\times$overlap factorial study (Section 4.5, 20 cells $\times$ 2 models $\times$ 50 documents = 2,000 extraction calls) cost $6.80.

I Licenses for Existing Assets

The synthetic datasets (Paystubs and W-2 Forms) are author-generated, while all other datasets are publicly available under permissive licenses (see Table 10).

Table 10: Datasets used in this study, their licenses, and sources. All third-party datasets are cited and accompanied by their stated licenses or terms of availability, where available. Synthetic datasets (Paystubs and W-2 Forms) are author-generated and released under CC-BY-4.0.

Asset	License	Source
SROIE	CC-BY-4.0	ICDAR 2019 Competition, no formal license declared by organizers
CORD v2	CC-BY-4.0	CORD v2 (100 test receipts, 93 qualifying) via naver-clova-ix/cord-v2 on HuggingFace
LEDGAR (via LexGLUE)	CC-BY-4.0	Accessed via LexGLUE 'coastalcph/lex_glue'), 50 bundles
VRDU	Apache-2.0	Google Research repo: Registration + Ad-Buy Domains
CUAD	CC-BY-4.0	CUAD v1, 50 bundles, via 'TheAtticusProject/cuad'
GST Invoices	CC-BY-4.0	Roboflow Universe Tax Invoice dataset, 100 of 773 images
Synthetic Paystubs	Author-generated	200 documents, 14 concepts
Synthetic W-2 Forms	Author-generated	50 forms, 44 concepts

Table 11 lists the two proprietary LLMs evaluated in this paper. Both providers permit research use under their respective terms of service.

Table 11: Proprietary API models and their providers.

Model	Provider	Terms
Claude Haiku 4.5 (claude-haiku-4-5-20251001)	Anthropic	Anthropic API Terms of Service and Usage Policy
GPT-4o-mini (gpt-4o-mini-2024-07-18)	OpenAI	OpenAI Terms of Use and Usage Policies

Details of the software dependencies and their licenses are provided in Table 12.

Table 12: Software dependencies and licenses.

Library	License	URL
Tesseract OCR	Apache-2.0	github.com/tesseract-ocr/tesseract
HuggingFace Datasets	Apache-2.0	github.com/huggingface/datasets
hindsight-eval	MIT	Released with this paper

J Extended Results Tables

J.1 Per-family misbinding breakdown for W-2 domain.

Table 13 shows the distribution of Haiku's 194 misbindings across ontology families (family of the ground-truth concept, regardless of which family it was bound to).

Table 13: Per-family misbinding count on W-2 forms (Claude Haiku 4.5, 50 forms).

Family	Misbindings	% of total
State/local	119	61%
Identity	31	16%
Compensation	23	12%
Box 12 (benefits codes)	15	8%
Withholding	6	3%
Other	0	0%

J.2 Top confusion pairs on W-2.

Table 14 lists the most frequently confused concept pairs from W-2, as referenced in the main text.

Table 14: Top concept confusion pairs on W-2 forms (Claude Haiku 4.5).

Ground Truth Concept	Predicted As	Count	Family
`state_2_income_tax`	`state_1_income_tax`	17	state_local
`state_2`	`state_1`	14	state_local
`state_2_wages`	`state_1_wages`	14	state_local
`employee_ssn`	`employer_ein`	18	identity
`social_security_tips`	`medicare_wages`	16	compensation
`local_1_income_tax`	`state_1_income_tax`	9	state_local

K Baseline Validation: Excluded Datasets

We considered several additional datasets during benchmark construction and excluded them for the following reasons:

- **FUNSD** (Jaume et al., 2019): its four coarse entity labels (*question, answer, header, other*) provide no concept-binding challenge, as the categories are trivially distinct.
- **DocVQA** (Mathew et al., 2021): the question-answering format evaluates one field at a time, sidestepping the multi-field competition that creates misbinding.
- **Kleister NDA** (Stanislawek et al., 2021): its field set (party names, dates, terms) has low semantic overlap; inclusion would not contribute to the semantic-overlap axis.
- **DeepForm** (Svetlichnaya, 2020): manual inspection revealed inconsistent annotation quality on campaign finance forms, which would confound misbinding measurements.
- **DocILE** (Simsa et al., 2023): while promising for its 55-class annotation granularity, access restrictions at the time of our experiments prevented inclusion. It remains a candidate for future extensions.

These exclusions are documented for transparency and to inform future benchmark extensions.

L CBA-SOFT DIAGNOSTIC: WITHIN-FAMILY CONFUSION ANALYSIS

L.1 CBA-SOFT AS AN ONTOLOGY-DIAGNOSTIC HEURISTIC

The gap between CBA-strict and CBA-soft is a diagnostic heuristic, not a calibrated measure of real-world severity: it equals half the fraction of recovered items whose misbinding stays within the same semantic family under the domain's *assumed* concept hierarchy (these receive partial, 0.5, rather than zero credit), i.e., $0.5 \times$ (within-family share of misbindings) $\times$ $(1 -$ CBA-strict$)$. At a fixed misbinding rate, a wider gap means more of the observed misbindings are within-family confusions under that hierarchy (e.g., `social_security_wages` confused with `medicare_wages`, both in the wages family); a narrower gap means a larger share are cross-family. On W-2 forms, CBA-soft exceeds CBA-strict by 4.5–11.0 percentage points (conditional on recovery: 86.2% vs. 75.3% for Haiku, 93.7% vs. 89.2% for GPT-4o-mini), consistent with the 77–89% within-family confusion rate reported in Section 4.2; most errors involve related fields such as state/local reporting pairs or SSN/EIN identifiers. On CUAD, the gap is 9.9–12.9 percentage points (84.4% vs. 74.5% for Haiku, 79.6% vs. 66.7% for GPT-4o-mini), and on LEDGAR, 7.6–9.8 percentage points (85.7% vs. 78.1% for Haiku, 84.9% vs. 75.1% for GPT-4o-mini), indicating a substantial share of within-family confusions receiving partial credit under each domain's ontology. We do *not* treat this gap as evidence of greater or lesser real-world deployment risk: whether a within-family confusion is actually low-stakes depends on downstream business impact, which the ontology's family groupings do not by themselves encode, and on researcher choices made when constructing the ontology (Section 3.4). Domain-specific severity calibration, in which family weights are tuned to reflect downstream business impact (e.g., heavier penalties on legal-clause inversions than on receipt-amount swaps) rather than assumed from the ontology structure, is left to future work.

L.2 SOFT–STRICT GAP BY DOMAIN

CBA-soft assigns partial credit (0.5) to within-family misbindings, reflecting the intuition that confusing two wage fields is *typically* less severe than confusing a wage with an identifier. The gap between CBA-soft and CBA-strict, therefore, equals $0.5 \times$ (within-family share of misbindings) $\times$ $(1 -$ CBA-strict$)$ under the domain's assumed ontology; it is a diagnostic heuristic tied to that ontology's construction (Section 3.4), not an independently calibrated measure of deployment severity.

Table 15 shows the soft-strict gap for high-misbinding domains, computed conditionally on recovery (CBA_{cond}-strict vs. its soft-credit counterpart) from the results of Table 2.

Table 15: CBA_{cond}-strict vs. CBA_{cond}-soft and the soft-strict gap.

Domain (Model)	CBA_{cond}**-strict**	CBA_{cond}**-soft**	**Soft−Strict**
W-2 (Haiku)	75.3%	86.2%	11.0pp
W-2 (GPT-4o-mini)	89.2%	93.7%	4.5pp
LEDGAR (Haiku)	78.1%	85.7%	7.6pp
LEDGAR (GPT-4o-mini)	75.1%	84.9%	9.8pp
CUAD (Haiku)	74.5%	84.4%	9.9pp
CUAD (GPT-4o-mini)	66.7%	79.6%	12.9pp

On W-2, the soft-strict gap ranges 4.5–11.0 percentage points; on LEDGAR, 7.6–9.8 points; on CUAD, 9.9–12.9 points, the largest of the three, indicating the greatest share of within-family confusions receiving partial credit under the CUAD ontology. This supports the use of CBA-soft as a diagnostic of ontology-relative confusion structure: at a given misbinding rate, a wide soft-strict gap signals that most errors are within-family under the assumed hierarchy. In contrast, a narrow gap indicates a more even mix. It should not, on its own, be read as a severity or deployment-risk ranking across domains, since the underlying family groupings reflect researchers' judgment at the time of ontology construction (Section 3.4) rather than measured business impact.

Future calibration of CBA-soft's partial-credit weights with domain-specific severity scores, as discussed in the main text, would further refine this diagnostic by replacing the ontology-relative within- and cross-family split with weights grounded in actual downstream impact.

M ADDITIONAL METRICS: PRECISION, DUPLICATES, AND JOINT F1

The metrics in Section 3.1 (Rec_{val}, CBA_{cond}, ACC_{joint}, Δ) are all recall-oriented: their denominator is $|GT|$, so a system that over-generates predictions (emitting extra, spurious concept-value pairs not present in the document) cannot be penalized by any of them. We define three additional metrics, computed from the same value-first bipartite matching (Algorithm 1), to close this gap.

Precision. The fraction of predicted items that correspond to a true ground-truth item, matched jointly on concept and value:

$$\text{Precision} = \frac{\left|\{(c'_j, v'_j) \in P : \exists (c_i, v_i) \in GT \text{ s.t. } \text{norm}(v'_j) = \text{norm}(v_i) \wedge c'_j = c_i\}\right|}{|P|}. \tag{7}$$

A model that hedges by emitting multiple candidate concepts per value, or extracting values not present in the document, will show high Rec_{val} but low precision; the joint metrics of Section 3.1 alone cannot reveal this.

Duplicate extraction rate. The fraction of predictions that repeat a $\text{norm}(v'_j)$ value already claimed by a higher-priority prediction in the same document (i.e., predictions the Hungarian assignment left unmatched because a lower-cost match to the same ground-truth item already exists):

$$\text{DupRate} = \frac{\left|\{(c'_j, v'_j) \in P : \text{norm}(v'_j) \text{ already matched by another } (c'_k, v'_k) \in P\}\right|}{|P|}. \tag{8}$$

A high duplicate rate indicates a model is emitting the same value under multiple candidate concept keys to game a recall-only baseline; Rec_{val} is invariant to this behavior by construction (Section 3.1), which is exactly why it must always be reported alongside precision, not as a replacement for it.

Joint F1 ($F1_{joint}$). The harmonic mean of ACC_{joint} (Equation 3, playing the role of joint recall) and Precision:

$$F1_{joint} = \frac{2 \cdot ACC_{joint} \cdot \text{Precision}}{ACC_{joint} + \text{Precision}}. \tag{9}$$

$F1_{joint}$ is the schema-aware key-value F1 referenced in Section 3.2: it is the single scalar that a standard, precision-aware structured-extraction evaluation would report, and CBA's contribution is to decompose disagreements between $F1_{joint}$ and Rec_{val} into their extraction- and binding-side causes rather than to replace $F1_{joint}$ as the headline number.